Operationalization of interactive Multilingual Access Gateways (iMAGs) in the Traouiero project

Christian Boitet, Valérie Bellynck, Achille Falaise, Hong-Thai Nguyen

Operationalization of interactive Multilingual Access Gateways (iMAGs) in the Traouiero project

French version included, obtained by MT + postediting with an iMAG, demo possible

Christian BOITET, Valérie BELLYNCK, Achille FALAISE, NGUYEN Hong-Thai

{ Christian.Boitet, Valerie.Bellynck, Achille Falaise, Hong-Thai.Nguyen }@imag.fr

GETALP, LIG-campus (UJF, CNRS, INPG, INRIA, UPMF)
BP 53, 38041 GRENOBLE cedex 9, France

ASLIB-2011, London, 18-20/11/2011

Abstract

We will explain and demonstrate iMAGs (interactive Multilingual Access Gateways), in particular on a scientific laboratory web site and on the Greater Grenoble (La Métro) web site. **This bilingual presentation has been obtained using an iMAG.**

Keywords

Interactive translation gateway, iMAG, MT post-editing, collaborative translation

Presentation

This presentation is an adaptation and update of an article presented as a demonstration only to TALN-2010. The names of the files have been kept the same, although their contents are slightly different. The iMAG concept has been proposed by Ch. Boitet and V. Bellynck in 2006 (Boitet & al. 2008, Boitet & al. 2005), and reached prototype status in November 2008, with a first demonstration on the LIG laboratory Web site. It has been adapted to the DSR (Digital Silk Road) Web site in April 2009, and then to more than 50 other Web sites. These first prototypes are extensions of the SECTra_w (Huynh & al. 2008) online translation corpora support system. Since the beginning of 2011, we are operationalizing this software with a view to deploy it as a multilingual access infrastructure, in the context of the French ANR (National Agency for Research) Traouiero "emergence" project.

An iMAG is an interactive Multilingual Access Gateway very much like Google Translate at first sight: one gives it a URL (starting Web site) and an access language and then navigates in that access language. When the cursor hovers over a segment (usually a sentence or a title), a palette shows the source segment and proposes to contribute by correcting the target segment, in effect post-editing an MT result. With Google Translate, the page does not change after contribution, and if another page contains the same segment, its translation is still the rough MT result, not the polished post-edited version. The more recent Google Translation Toolkit enables one to MT-translate and then post-edit online full Web pages from sites such as Wikipedia, but again the corrected segments don't appear when one later browses the Wikipedia page in the access language.

By contrast, an iMAG is dedicated to an elected Web site, or rather to the elected sublanguage defined by one or more URLs and their textual content. It contains a translation memory (TM) and a specific, preterminological dictionary (pTD), both dedicated to the elected sublanguage. Segments are pretranslated not by a unique MT system, but by a (selectable) set of MT systems. Systran and Google are mainly used now, but specialized systems developed from the postedited part of the TM, and based on Moses, will be also used in the future.

The powerful online contributive platforms SECTra_w and PIVAX (Nguyen & al. 2007) are used to support the TMs and pTDs. Translated pages are built with the best segment translations available so far. While reading a translated page, it is possible not only to contribute to the segment under the cursor, but also to seamlessly switch to SECTra_w online post-editing environment, equipped with proactive dictionary help and good filtering and search-and-replace functions, and then back to the reading context.

A translation relay is being implemented to define the iMAGs or other translation gateways used by an elected Web site, select and parameterize the MT systems and translation routes used for various language pairs, and manage users, groups, projects (some contributions may be organized, other opportunistic), and access rights. Finally, MT systems tailored to the selected sublanguage can be built (by combinations of empirical and expert methods) from the TM and the pTD dedicated to a given elected Web site. That approach will inherently raise the linguistic and terminological quality of the MT results, hopefully converting them from rough into raw translations. The demonstration will use some iMAGs created by the AXiMAG startup for various Web sites, such as those of the LIG lab (http://service.aximag.fr:8180/xwiki/bin/view/imag/liglab) and of La Metro (Greater Grenoble) web site (http://service.aximag.fr:8180/xwiki/bin/view/imag/lametro), where access in Chinese and English was enabled in 2010 for the Shanghai Expo.

In this written presentation, we will apply the iMAG technique to the presentation itself. It was first written in English under Word, in file TALN-2010-Demo-iMAG-v2_en.rtf. We then saved it in html and put the result (TALN-2010-Demo-iMAG-v2_en.htm) online (www-clips.imag.fr/geta/User/christian.boitet/iMAGs-tests/en). We accessed it through the corresponding iMAG (http://service.aximag.fr:8180/xwiki/bin/view/imag/xan-en). Choosing French as access language, and tuning some parameters, we got the following view.

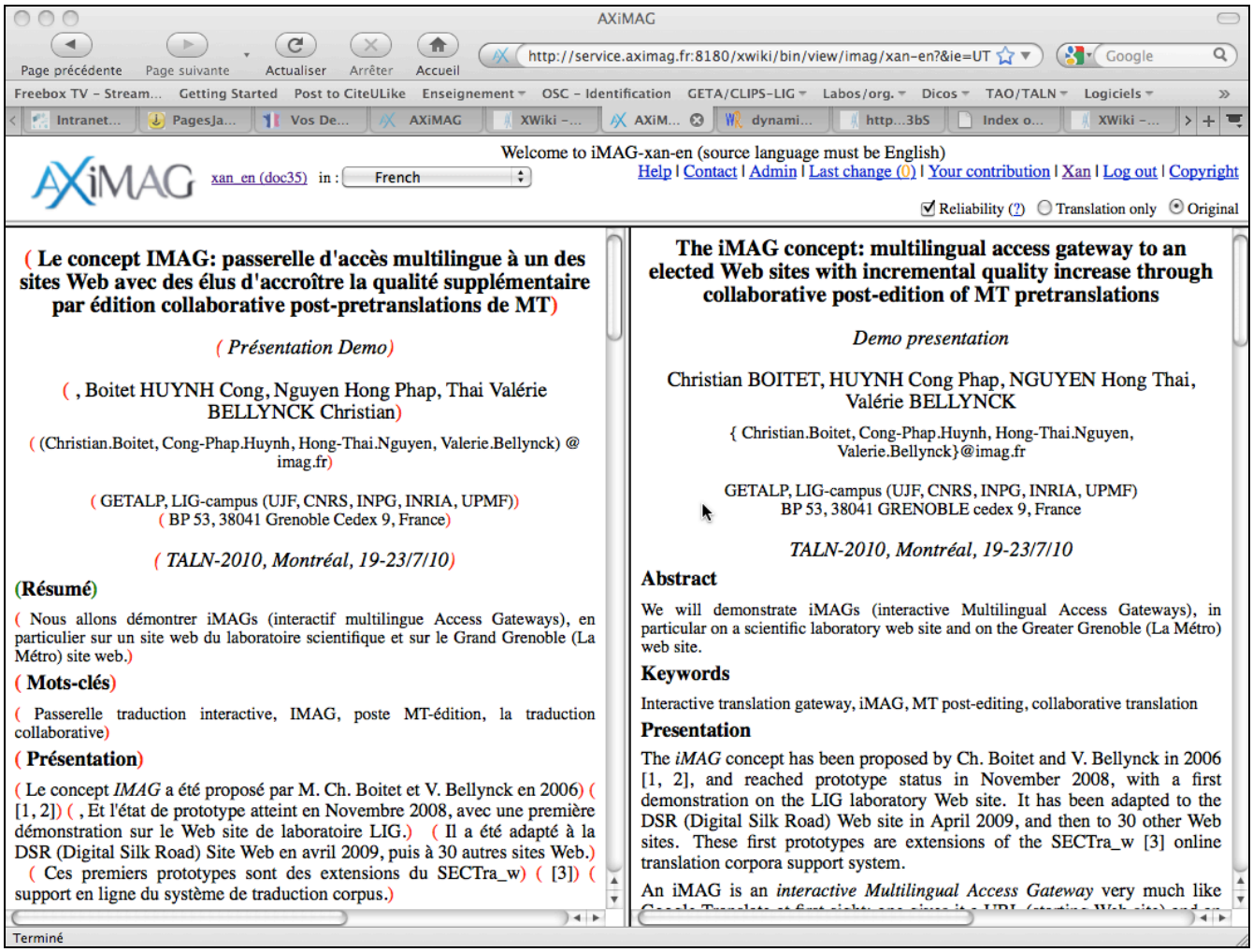

Figure 1 : screen of of iMAG after MT, parallel presentation — écran d'une iMAG après TA, présentation parallèle

Note the colored curly brackets enclosing the French segments. They appear on demand ("reliability" checkbox). Red indicates an MT output, green an postedition by a certified translator, and orange a postedition by somebody knowing the two languages to a certain level, and usually contributing on an occasional and volunteer basis. After postediting some segments directly on the Web page, we obtain the following state (notice some green brackets).

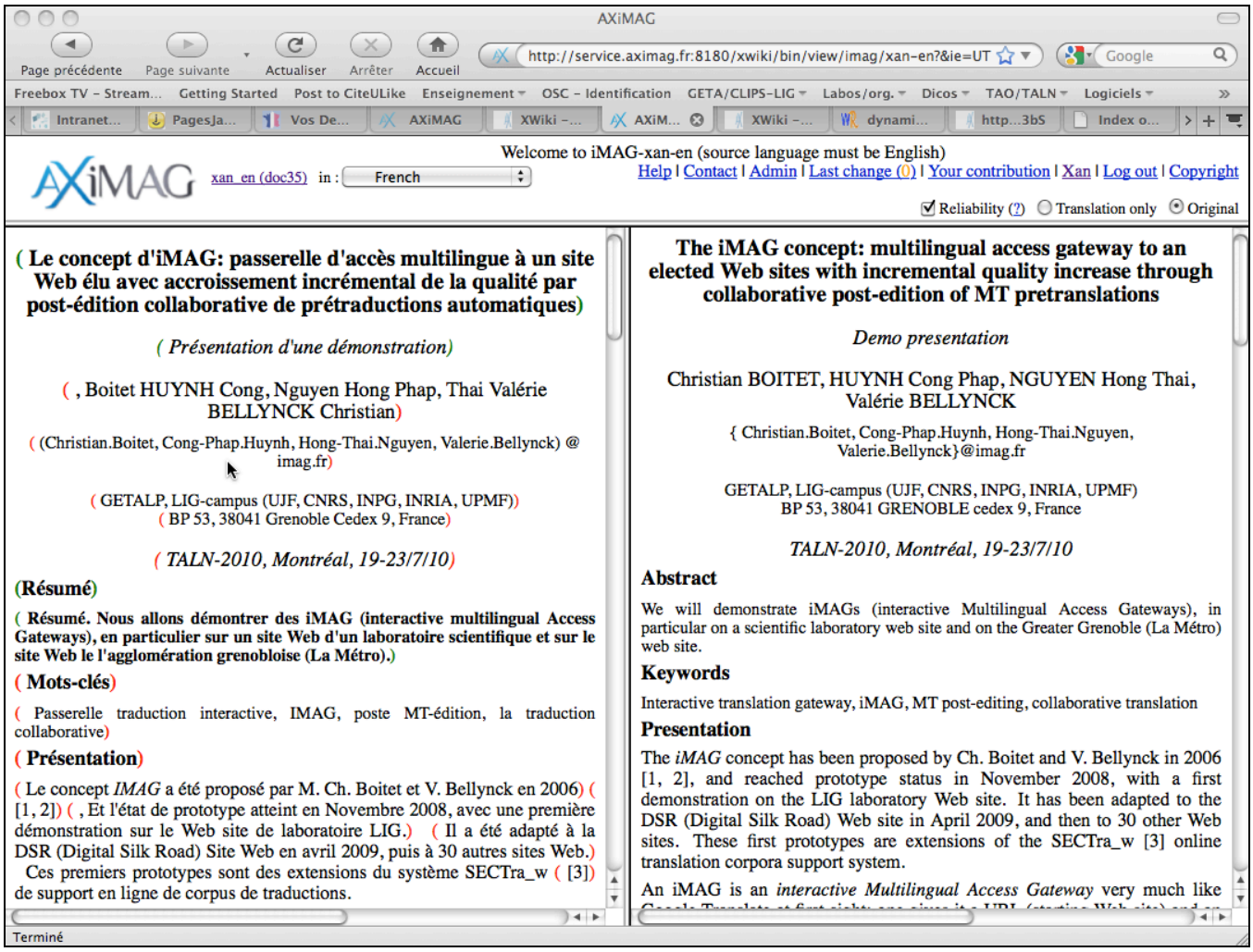

Figure 2 : screen of an iMAG after direct postediting of some segments on a Web page — écran d'une iMAG après postédition directe de quelques segments sur une page Web

Having logged on as a "certified" contributor, we could continue in "advanced mode", switching to the SECTra_w online postediting environment, under which postediting is much faster (specialized lexical help is not available until the corpus of postedited segments is large enough, so that it is not shown here).

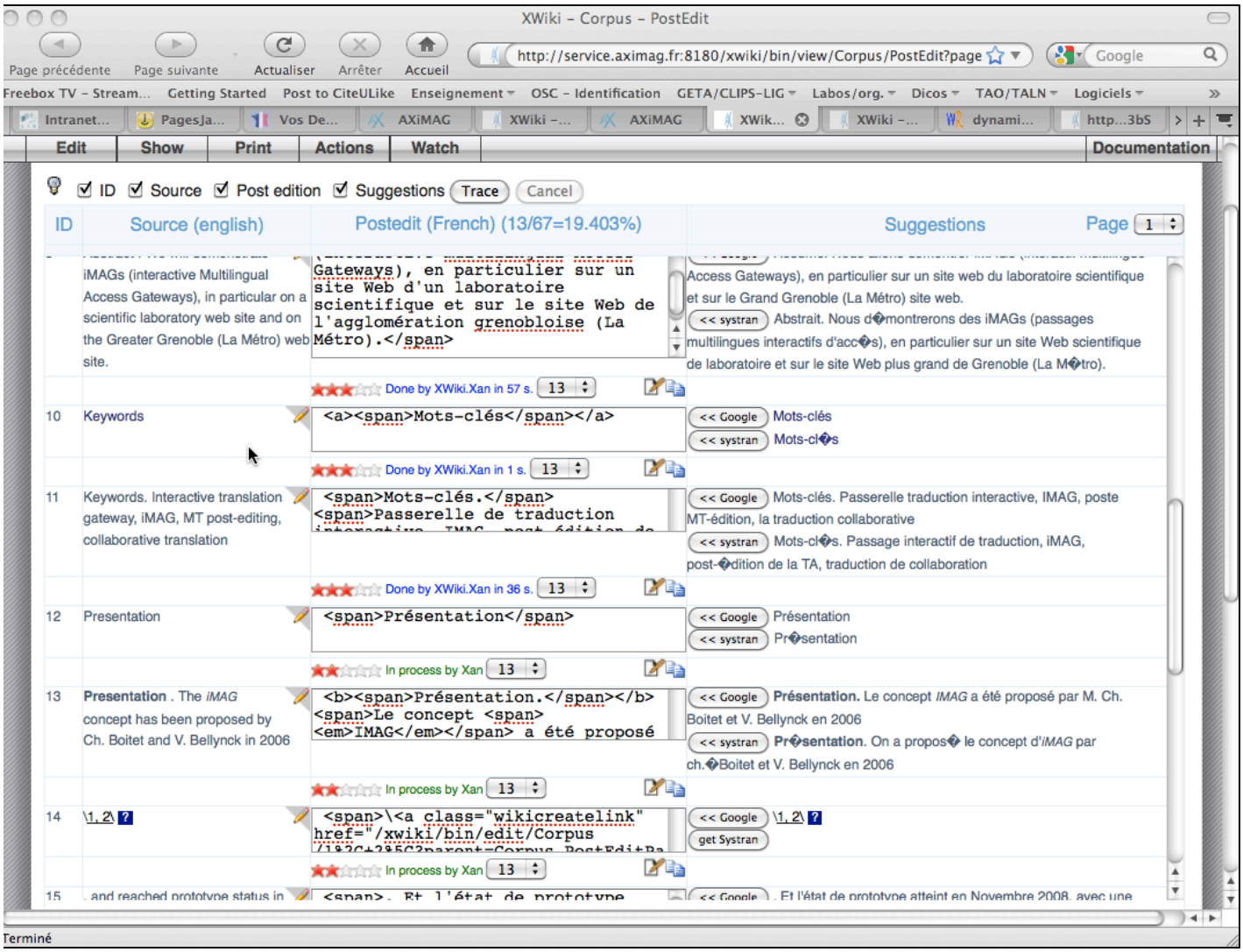

Figure 3 : screen of an iMAG after indirect postediting under SECTra_w of other segments — écran d'une iMAG après postédition indirecte sous SECTra_w d'autres segments

After having postedited the whole presentation, we extracted the current French version from SECtra_w. After a simple manipulation (to be automated in the future), the text was obtained in a bilingual table:

By contrast, an iMAG is dedicated to an elected Web site, or rather to the elected sublanguage defined by one or more URLs and their textual content.	En revanche, une IMAG est dédiée à un site Web élu, ou plutôt au sous-langage élu défini par une ou plusieurs URL et leur contenu textuel.
It contains a translation memory (TM) and a specific, preterminological dictionary (pTD), both dedicated to the elected sublanguage.	Elle contient une mémoire de traductions (MT) et un dictionnaire spécifique preterminologique (pTD), les deux dédiés au sous-langage élu.
Segments are pretranslated not by a unique MT system, but by a (selectable) set of MT systems. Systran and Google are mainly used now, but specialized systems developed from the postedited part of the TM will be also used in the future.	Les segments sont prétraduits non pas par un système de TA unique, mais par un ensemble (sélectionnable) de systèmes de TA. Systran et Google sont principalement utilisés aujourd'hui, mais des systèmes spécialisés développés à partir de la MT postéditée seront également utilisés dans l'avenir.

Figure 4 : excerpt of a bilingual aligned text extracted from an iMAG-translated target Web page — extrait d'un text bilingue aligné provenant d'une page Web traduite par une iMAG

After 15 more minutes of formatting, we got the French text that follows. References are factorized after the French version as the custom is not to translate them when translating between English and French.

Opérationnalisation de passerelles interactives d'accès multilingue (iMAG) dans le projet Traouiero

Version française incluse, obtenue par TA + postédition avec une iMAG, démo possible

Christian BOITET, Valérie BELLYNCK, Achille FALAISE, NGUYEN Hong-Thai

{ Christian.Boitet, Valerie.Bellynck, Achille Falaise, Hong-Thai.Nguyen }@imag.fr

ASLIB-2011, London, 18-20/11/2011

Résumé

Nous expliquerons et présenterons des iMAG (interactive multilingual Access Gateways), en particulier sur un site Web d'un laboratoire scientifique et sur le site Web de l'agglomération grenobloise (La Métro). Cette présentation bilingue a été obtenue en utilisant une iMAG.

Mots-clés

Passerelle de traduction interactive, IMAG, post-édition de TA, traduction collaborative

Présentation

Cette présentation est une adaptation et une mise à jour d'un article présenté uniquement comme une demonstration à TALN-2010. Les noms des fichiers sont les mêmes, bien que leurs contenus soient un peu différents. Le concept d'IMAG a été proposé par M. Ch. Boitet et V. Bellynck en 2006 (Boitet & al. 2008, Boitet & al. 2005), et a atteint l'état de prototype en novembre 2008, avec une première démonstration sur le site Web du laboratoire LIG. Il a été adapté au site Web DSR (Digital Silk Road) en avril 2009, puis à plus de 50 autres sites Web. Ces premiers prototypes sont des extensions du système SECTra_w (Huynh & al. 2008) de support en ligne de corpus de traductions.

Depuis début 2011, nous sommes en train *d'opérationnaliser* the logiciel en vue de le déployer comme une *infrastructure d'accès multilingue,* dans le contexte du projet "emergence" Traouiero de l'ANR (Agence Nationale de la Recherche).

Une iMAG est une *passerelle interactive d'accès multilingue* (interactive Multilingual Access Gateway), ressemblant beaucoup à Google Translate, à première vue: on donne une URL (site Web de départ) et une langue d'accès, et on navigue ensuite dans cette langue d'accès. Lorsque le curseur passe sur un segment (le plus souvent une phrase ou un titre), une palette montre le segment source et propose de contribuer en corrigeant le segment cible, en fait en post-éditant un résultat de TA. Avec Google Translate, la page ne change pas après la contribution, et si une autre page contient le même segment, sa traduction est toujours le résultat de TA grossière, pas la version polie post-éditée. La boîte à outils de traduction plus récente Google Translation Toolkit permet de traduire par TA et ensuite de poster en ligne des pages Web complètes tirées de sites tels que Wikipedia, mais de nouveau, les segments corrigés n'apparaissent pas quand on regarde plus tard la page de Wikipedia dans la langue d'accès.

En revanche, une IMAG est *dédiée à un site Web élu,* ou plutôt au *sous-langage élu* défini par une ou plusieurs URL et leur contenu textuel. Elle contient une *mémoire de traductions* (MT) et un *dictionnaire spécifique preterminologique* (pTD), les deux dédiés au sous-langage élu. Les segments sont *prétraduits* non pas par un système de TA unique, mais *par un ensemble (sélectionnable)* de systèmes de TA. Systran et Google sont principalement utilisés aujourd'hui, mais des systèmes spécialisés développés à partir de la MT postéditée, et bases sur Moses, seront également utilisés dans l'avenir.

Les plates-formes contributives puissantes *SECTra_w* et *PIVAX* (Nguyen & al. 2007) sont utilisées pour supporter les MT et les pTD. Les pages traduites sont construites avec les meilleures traductions des segments disponibles à ce jour. Pendant la lecture d'une page traduite, il est possible non seulement de contribuer au segment sous le curseur, mais aussi de passer de façon transparente sous l'environnement de postédition en ligne de SECTra_w, muni d'une *aide dictionnairique proactive* et de bonnes fonctions de filtrage et de recherche-remplacement, et ensuite de revenir dans le contexte de lecture.

Un *relais de traduction* est en cours d'implémentation pour définir les iMAG ou d'autres passerelles de traduction utilisées par un site Web élu, pour sélectionner et paramétrer les systèmes de TA et les itinéraires utilisés pour la traduction des différentes paires de langues, et pour gérer les utilisateurs, les groupes, les projets (certaines contributions peuvent être organisées, d'autres opportunistes), et les droits d'accès. Enfin, des *systèmes de TA faits sur mesure* pour le sous-langage sélectionné *peuvent être construits* (par la combinaison de méthodes empiriques et expertes) à partir de la MT et la du pTD dédiés à un site Web élu donné. Cette approche accroîtra intrinsèquement la qualité linguistique et terminologique des résultats de TA, en les transformant de traductions grossières en traductions brutes. La démonstration utilisera des iMAG créées par la jeune pousse AXiMAG pour divers sites Web, tels que ceux du laboratoire LIG (http://service.aximag.fr:8180/xwiki/bin/view/imag/liglab) et du site Web de La Métro (agglomération grenobloise)

(http://service.aximag.fr:8180/xwiki/bin/view/imag/lametro), où l'accès en chinois et en anglais a été activé en 2010 pour l'Expo de Shanghai.
Dans cette présentation écrite, nous allons appliquer la technique iMAG à la présentation elle-même. Elle a d'abord été écrite en anglais sous Word, dans le fichier TALN-2010-Demo-IMAG-v2_en.rtf. Nous l'avons ensuite sauvée en format html, et avons mis le résultat (TALN-2010-Demo-IMAG-v2_en.htm) en ligne (www-clips.imag.fr/geta/User/christian.boitet/iMAGs-tests/en). Nous y avons accédé par l'intermédiaire de l'iMAG correspondante (http://service.aximag.fr:8180/xwiki/bin/view/imag/xan-en).
Après avoir choisi le français comme langue d'accès, et réglé certains paramètres, nous avons obtenu la vue suivante — voir Figure 1.
Notez les accolades de couleur entourant les segments français. Elles apparaissent à la demande (case à cocher "fiabilité"). Le rouge indique une sortie de TA, le vert une postédition par un traducteur certifié, et l'orange une postédition par quelqu'un connaissant les deux langues à un certain niveau, et de contribuant généralement de façon occasionnelle et à titre bénévole. Après avoir postédité certains segments directement sur la page Web, on obtient l'état suivant (notez certains crochets verts) — voir Figure 2.
Après nous être connecté en tant que contributeur "certifié", nous avons pu continuer en mode "avancé", en passant sous l'environnement de postédition en ligne de SECTra_w, sous lequel la postédition est beaucoup plus rapide (l'aide lexicale spécialisée n'est pas disponible jusqu'à ce que le corpus de segments postedités soit assez grand, donc elle n'est pas montrée ici) — voir Figure 3.
Après avoir postédité toute la présentation, nous avons extrait la version française courante de SECtra_w,. Après une manipulation simple (qui sera automatisée dans le futur), le texte a été dans un tableau bilingue :

By contrast, an iMAG is dedicated to an elected Web site, or rather to the elected sublanguage defined by one or more URLs and their textual content.	En revanche, une IMAG est dédiée à un site Web élu, ou plutôt au sous-langage élu défini par une ou plusieurs URL et leur contenu textuel.
It contains a translation memory (TM) and a specific, preterminological dictionary (pTD), both dedicated to the elected sublanguage.	Elle contient une mémoire de traductions (MT) et un dictionnaire spécifique preterminologique (pTD), les deux dédiés au sous-langage élu.
Segments are pretranslated not by a unique MT system, but by a (selectable) set of MT systems. Systran and Google are mainly used now, but specialized systems developed from the postedited part of the TM will be also used in the future.	Les segments sont prétraduits non pas par un système de TA unique, mais par un ensemble (sélectionnable) de systèmes de TA. Systran et Google sont principalement utilisés aujourd'hui, mais des systèmes spécialisés développés à partir de la MT postéditée seront également utilisés dans l'avenir.

Figure 5 : partie d'un texte bilingue aligné extrait d'une page Web cible traduite par une iMAG

Les références ont été factorisées après la version en français car la coutume est de ne pas traduire les références entre anglais et français.

References

BOITET, C., Y. BEY et K. KAGEURA (2005) Main research issues in building web services for mutualized, non-commercial translation. *Proc. SNLP-05,* Bangkok, Chulalongkorn university, 13 p.

BOITET, C., V. BELLYNCK, M. MANGEOT et C. RAMISCH (2008) Towards Higher Quality Internal and Outside Multilingualization of Web Sites. *Proc. ONII-08 (Summer Workshop on Ontology, NLP, Personalization and IE/IR),* Mumbai, Centre for Indian Language Technology (CFILT), IITB (Indian Institute of Technology Bombay), 9 p.

HUYNH, C.-P., C. BOITET et H. BLANCHON (2008) SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08,* Marrakech, 8 p.

NGUYEN, H.-T., C. BOITET et G. SÉRASSET (2007) PIVAX, an online contributive lexical database for heterogeneous MT systems using a lexical pivot. *Proc. SNLP-07,* Pattaya, Kasetsart University, 6 p.

Appendix: screenshots of two other iMAGs / images d'écran de deux autres iMAG

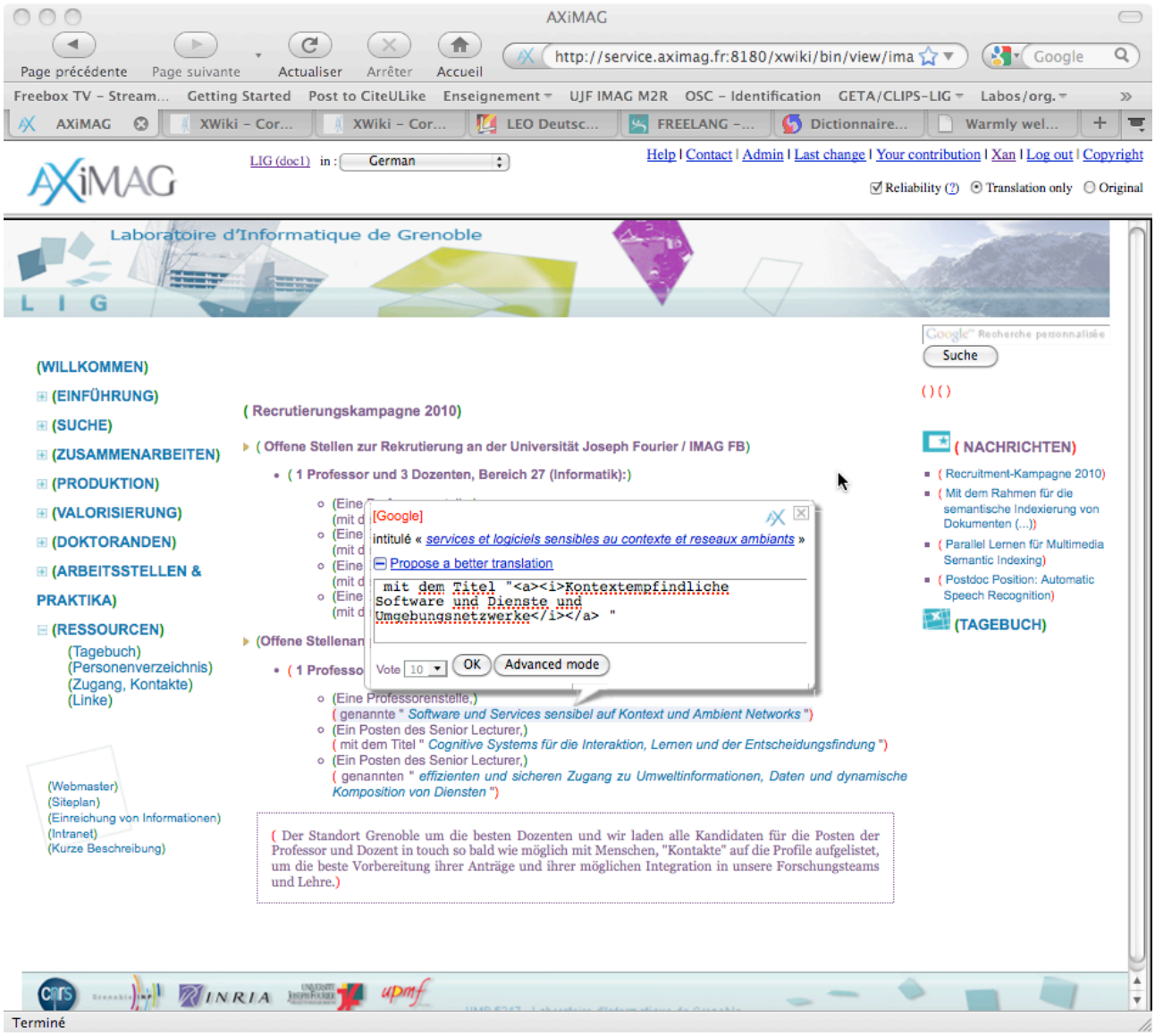

Figure 6: demonstration on the LIG
(Grenoble Informatics Laboratory — Laboratoire d'Informatique de Grenoble) web site —
demonstration sur le site Web du LIG

Figure 7: demonstration on the La Métro (Greater Grenoble) web site

Short CV (Christian Boitet)

Ch. Boitet is professor of computer science at Université Joseph Fourier (Grenoble 1), where he has taught algorithmics, compiler construction, formal languages & automata, elementary logic, formal systems, and natural language processing. He is one of the authors of Ariane-G5, GETA's generator of MT systems. He has presented communications in many national and international conferences and published in various journals and books. He has also edited a book dedicated to the presentation of Pr. Vauquois' scientific work, as well as several international conference proceedings. He has been in charge of several research contracts aiming at reaching the operational stage and the industrial stage. He has also been involved in and/or in charge of GETA's participation in several cooperative research efforts. He is a member of ICCL and has been one of the organizers of COLING-92. In 1998, he was Programme Chair of COLING-ACL'98. He is and has been a regular reviewer for several journals and conferences, and has been in the programme committees of many congresses. His current interests include personal dialogue-based MT for monolingual authors (GETA's LIDIA project, international UNL project), speech translation (CSTAR project), machine helps to translators and interpreters, integration of speech processing inspired techniques in MT, multilingual lexical data bases (Papillon project), and specialized languages and environments for lingware engineering and linguistic research (Ariane-Y project).

He introduced several innovations in various aspects of MT, such as multitarget pidgin translation with limited reordering (Russian to French & English, TAUM, 1972), contextual rewriting rules in transformational systems for decorated trees (ROBRA, COLING-78), vertical parallelism and guarded recursion and iteration in ROBRA, a formalization of structured string-tree correspondences (SSTCs, with Zarin Y., COLING-88) later used as the basis of the SiSTeC-ebmt systems (USM & MMU, Malaysia), the whiteboard architecture for heterogeneous speech translation systems (with M. Seligman, 1993), the use of text processors as pseudo-syntactic editors to develop multitarget dictionaries (French-English-Malay or FEM dictionary, 1996), the LIDIA architecture based on Interactive Disambiguation by Auhors and the concept of self-explaining documents (SED, with H. Blanchon, MT15YO, 1994), the concept of metaenvironment for MT, the definition of heterogeneous MT systems (with Nguyen H.T., 2009), several concepts to enable rigorous processing of aligned translation corpora (with Huynh C.P., 2010), such as multilingualized segments, pseudodocuments, metadocuments, multiple and recursive segmentation, and techniques to implement incrementally modifiable MT systems for unlimited translation units (with JC. Durand, 2011). More general contributions are the distinctions between linguistic, computational and operational architectures of MT systems, and between strong and weak translation problems (with A. Malik, 2010).